数字图书馆数值知识元检索系统设计*

黄容[1] 何杨煜琪[1] 王忠义[1] 李春雅[2]
[1] 华中师范大学信息管理学院 武汉 430079 [2] 南通理工学院商学院 南通 226002

摘要：[目的/意义]为满足数字图书馆用户对数值知识的个性化检索需求，向其提供细粒度的知识服务。[方法/过程]基于对数值知识元的深入分析，提出数字图书馆数值知识元识别、抽取、索引与检索方法，并构建一个面向数值知识元的检索系统。[结果/结论]通过实例分析验证基于数值知识元的细粒度知识服务能够在一定程度上提高检索和利用数值知识的效率和用户满意度。

关键词：数值知识元 知识元识别 知识元标引 知识元检索
分类号：G250
DOI：10.13266/j.issn.0252-3116.2018.14.015

1 引言

21世纪以来，随着以信息技术为代表的科学技术的发展，人们已然进入一个信息爆炸的时代，知识服务的出现给数字图书馆的工作带来了许多挑战[1]，现有的数字图书馆的数字产品大都采用基于主题词的模式进行资源的组织和服务，知识服务的基本单位通常还是文献，无法针对具体问题向用户提供细粒度的知识服务[2-4]。比起文献级别的知识单位，人们更多地希望能够直接检索到自己感兴趣的知识点，这就要求数字图书馆将知识的控制单位逐渐由粗粒度的文献单元深化到细粒度的知识元单元[5]，实现从对知识载体和知识属性特征的管理到对知识内容本身的管理，也即变间接知识管理方式为直接知识管理方式[6]。知识元是不可再分割的具有完备知识表达的知识单位[7]。依据知识元内容的不同，可以将知识元划分为理论与方法知识元、数值知识元、事实知识元等多种类型[8]。其中，数值知识元是指以数值形式存在的，描述客观事物或者事件有关数值方面属性（如时间、长度、高度、重量、百分比、销售额、利润等）的知识单元[9]。数值知识元对于推动数值知识的利用，提高人们检索和利用数值知识的效率，帮助人们发现潜在的、隐含的数值知识关系等具有非常重要的意义。当前大多学者是从理论的角度对数值知识元进行了研究[10-14]，但如何更有效地从文本中抽取完整、准确的数值知识元，仍然需要进一步深入研究。为此，本研究以数字图书馆数字馆藏资源为研究对象，对数字图书馆数值知识元的识别、抽取、标引与检索进行研究，以期细化数字图书馆知识服务的粒度，提高数字图书馆知识服务效率。

2 数值知识元的识别与抽取

2.1 数值知识元的识别与抽取规则

从数字图书馆馆藏数字资源中识别数值知识元，首先应考虑知识资源的存在形式[15-16]。知识不仅储藏在传统的文献数据库中，还广泛分布在专利数据、行业标准、科技报告等特色资源库中。本研究的研究对象是数值知识元，数字馆藏资源中数值知识元的描述多以句子为单位，这种情况比较适合规则与模式识别方法。为此，数值知识元包括哪些类型，以及如何构建数值知识元的识别规则，是从数字馆藏资源中识别出数值知识元的关键。

数值知识元的识别是通过计算识别规则与知识元的匹配关系来实现的。判断特征标识之后的段落和句子是否具有包含规则标识描述的知识元的内容，若有，则特征标识就是向导信息，其后的具体内容就是知识元；否则特征标识就不被选中[17]。为从数字馆藏资源中归纳出数值知识元的识别规则，本研究首先对数值知识元进行划分。由于数值按照功能作用可以分为3

* 本文系教育部人文社会科学研究青年基金项目“数字图书馆馆藏资源多粒度层级主题分隔研究”（项目编号：16YJC870003）研究成果之一。
作者简介：黄容（ORCID：0000-0002-2791-0042），硕士研究生；何杨煜琪（ORCID：0000-0001-6247-2394），本科生；王忠义（ORCID：0000-0001-8945-783X），副教授，硕士生导师；李春雅（ORCID：0000-0001-7155-4658），讲师，博士，通讯作者，E-mail：10880945@qq.com。
收稿日期：2018-01-29 修回日期：2018-04-04 本文起止页码：125-132 本文责任编辑：王传清

类:基础数值、过程数值、结果数值,因此,本研究将数字馆藏资源中由这3类数值构成的数值知识元分为基础数值知识元、过程数值知识元、结果数值知识元3种类型。不同类型的数值知识元有着不同的描述方式,句子的结构和复杂度也有较大的差异,数值知识元的流程虽然可以通过有些数值辅助判定,但用句群或段落进行描述更加完整、准确。通过对数值知识元的类型剖析以及描述规则的构建,可以辅助识别出文本中数值知识元的位置,有助于后续的数值知识元抽取的实现。

为归纳出3种类型的数值知识元的识别规则,本研究采用文本分析法,首先从13门学科中的核心期刊选取20篇文献,共计260篇文献资源,通过对这些核心期刊论文进行分句,提取其中含有数值信息的完整句子,接着筛选、分析归纳出不同类型的数值知识元比较共性的表达方式,据此构建出描述规则如表1所示:

表1 数值知识元识别规则

类型	数值知识元结构	数值知识元识别规则	举例
基础数值知识元	时间+主体+Source+谓词+数值+单位+指标	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+主体+(从/在/以/选取…)+Source+(回收/收集/采集/发放/获取/选取/下载/提供/进行/得/为/有/是/达到/有/共计…)+数值+单位(如:个、篇、件、元等)+指标	截至2010年12月,从中国引文数据库下载了相关领域的100篇文章
	时间+主体+Source+谓词+指标+数值+单位	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+主体+(从/在/以/选取…)+Source+(回收/收集/采集/发放/获取/选取/下载/提供/进行/得/为/有/是/达到/有/共计…)+指标+数值+单位(如:个、篇、件、元等)	采集时间为2012年3月,对万方和CSSCI两大中文期刊数据库收集了论文数据共计4886篇
过程数值知识元	时间+主体+指标+谓词+数值+单位	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+主体+(最大值/最小值/权重/阈值/维度/临界值/相似值/…率/)+(达到/为/非/介于/处于/取/为/大于/等于/小于…)+数值+单位	2010年5月8日,各试件的位移延性系数均达到了3.0
	时间+数值+单位+主体+指标+谓词	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+在数值~数值+单位+范围内+主体+指标+谓词	截至2016年7月20日,在300–600℃范围内,热失重的速率增大。
结果数值知识元	时间+主体+指标+谓词+数值+单位	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+主体+(中/过/好/到/有/定/含/内)的((分别/均/仅)(认/设/定/成分/示/本/改/否)为/达到/仅有/下降/上升/提高到/大概为/最低为)+数值+单位	2013年,人才网站的查全率达到80.67%
	时间+数值+单位+主体+指标+谓词	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+数值+单位+主体+指标+谓词	截至2015年6月,56位核心作者的文献被引数超过了其他
	时间+主体+谓词+数值+单位+指标	(…年…月…日~/至…年…月…日)/(…年…月…日)/(截至…/截止…/日期/时间为…)+主体+(获得/得到/实现/取得)+数值+单位+指标	2015年10月8日,当地获得了政府3 000万元的补助

2.2 基于规则的数值知识元的识别与抽取

依据数值知识元结构与识别规则,本研究设计数值知识元的抽取方法,见图1。数值知识元抽取的基本流程包括:文本分句、分词及词性标注、句子过滤、数值知识元属性识别与抽取等步骤。由于基于内容分析法的数值知识元识别规则的提取已经在2.1节详细论述,因此,下文将详细论述其它各步骤的具体实现过程。

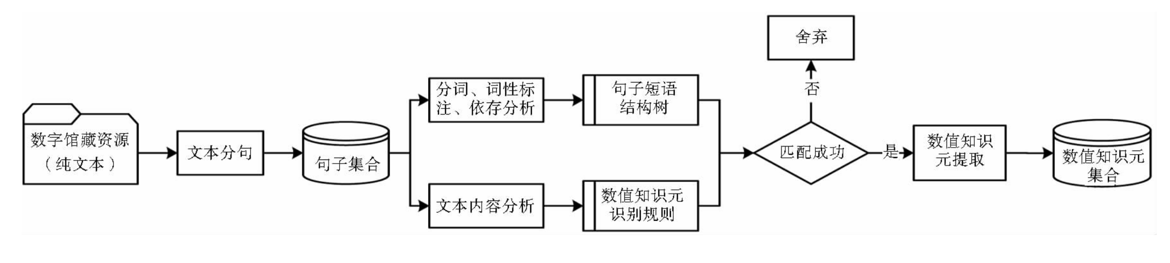


图1 数值知识元的抽取流程

2.2.1 文本分句 首先将选取的文献资源(PDF格式)转变为纯文本格式,去掉其中的不相关信息,例如文献目录、图片等,依据语句标识符(如。;?! 等)对文本进行分句。

2.2.2 分词及词性标注 以句子为单位,生成句子短语结构,并进行词性标注,对文本中的每个词选择一个最有可能的词类,包括名词、动词、数词、量词等;去除停用词,包括形容词、冠词等语义内容很少的词。

黄容，何杨煜琪，王忠义，等.数字图书馆数值知识元检索系统设计[J].图书情报工作,2018,62(14):125－132.

2.2.3　匹配　前文中数值知识元的识别规则定义了不同类型数值知识元语句中的线索词(如数值、单位等)以及这些线索词之间的组配结构。基于这些数值知识元识别规则,将其与分词后的语句进行匹配,定位到数值知识元所在句群或段落,从而达到提取符合这些规则的候选句子的目的,它们构成了数值知识元抽取的语句集合。

2.2.4　数值知识元提取　该步骤的主要任务是根据数值知识元结构,获取数值知识元的主体、指标、时间、谓词、数值、单位、来源等属性。各属性的具体抽取规则如下:

(1)主体的识别与抽取。数值知识元的主体主要包括地区、行业、机构3种类型,可借助行业词表、机构特征词表,地名专用切词词典等识别行业名称、中文机构名称、县级以上行政区域以及县以下地域等,其中,如果文本中没有指明地区,则为"中国"。

(2)指标的识别方法。数值知识元中的指标是指知识元所表达的数值信息主题,一般以名词为主,与数值或单位相邻组合成短语,可以采取中文自动分词技术以及词性标注,抽取出数值知识元中的指标,并建立指标库辅助数值知识元主体的抽取。

(3)时间的识别方法。在文本信息中,时间信息的表达方式复杂多样,不仅仅是简单的日期表示,还包括复合时间短语、段时间词等,例如"截止今年6月份""2015年10月5日上午"等。为了准确的识别这些复杂的时间表达形式,将时间的表达方式归纳为一般化的4种:时间(例如:九点四十分)、日期(例如:2017年4月12日)、时间词(例如:今年、上午)、段时间(例如:一个月、两年)。而在数值知识元的识别过程中,根据本文定义的时间信息的表达模式,可以通过抽取同一分句中距离指标最近且最新的时间来识别数值知识元的时间。对于时间不具体的知识元,直接删除不要。

(4)谓词的识别方法。根据数值知识元的表达模式,谓词一般处于数值或指标的前方,且词性为动词或者介词与动词的结合,如"比去年下降"。

(5)数值的识别方法。根据汉语用语习惯,数值信息可以分为3类:基数类数值,是指相对单纯的数字,包括整数、小数、分数等,例如五十、百分之五、六点五等;序数词,以某些基数词与"第"的组合方式为主,例如第五、第二等;特殊数词,是指用非基数词的汉字表示数量、程度或范围的形式,例如若干、大半等。

(6)单位的识别方法。单位是指与数值进行组合的量词,例如个、只、元等,可以采用有限自动机算法依据量词模式库匹配和抽取数值之后的量词。

(7)来源的识别方法。在文本分析时即可识别出该条信息出自哪篇文献,并抽取出其URL地址。

2.2.5　生成数值知识元　将结果进行存储,形成数值知识元库,以方便数值知识元的进一步共享与分析利用。

3　数值知识元的索引与检索

数值知识元是数字图书馆数值知识构建的基元,数值知识元索引与检索更是细粒度知识组织与服务的重要环节,对数值知识元的存储、检索与使用具有重要的意义。

3.1　数值知识元的描述架构

目前,国内对数值知识元描述架构的研究较少,学者们的研究成果也具有差异。本研究通过分析,提出更为一般化的数值知识元实体对象结构的描述框架,该框架从知识标识、知识描述、知识关系3个层面构建数值知识元的实体对象结构,具体如表2所示:

表2　数值知识元描述架构

结构层次	描述内容
标识组	数值知识元的名称
描述组	数值知识元的时间、主体、指标、谓词、数值、单位
关系组	数值知识元的来源

其中,知识标识组描述数值知识元在存储、利用方面的唯一标识,例如数值知识元的名称,是对该数值知识元内容的一种高度概述;知识描述组主要描述数值知识元的本质内容与内在属性,如数值知识元的主体、指标、数值、单位等;知识关系组主要描述数值知识元与资源间的关系,例如数值知识元的来源,可以链接到包含该知识元的载体,从而获得更加完整的相关信息。

3.2　数值知识元索引

数值知识元的索引工作是指对数值知识元所讨论的主题与相关属性(如数值、主体、指标等)构建索引,以确定其检索标识和指出其所在位置,实现快速准确检索的目标[18]。基于上文中提出的数值知识元的描述架构,本研究决定从知识标识、知识描述、知识关系3方面对数值知识元进行索引,现提出数值知识元的标引流程,见图2。数值知识元的索引流程主要包括:信息抽取模块、分词模块、特征提取模块和索引建立模块。

3.2.1　信息抽取　信息抽取模块的主要功能是从数值知识元库中抽取出相关信息,以便建立面向各数值知识元的索引。由于本研究在对数值知识元进行索引

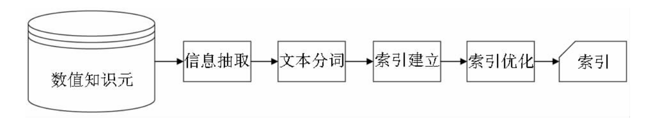


图 2　数值知识元的标引流程

时,借助 Lucene 这一高性能的搜索引擎架构,该索引架构以 Document 为索引的基本单位,而 Document 作为 Lucene 索引对象,又是由许多域(field)构成,因此,信息抽取的任务就转化为从数值知识元中抽取出组成 Document 的域的过程。由数值知识元的描述架构可知,数值知识元主要包括:名称、时间、主体、指标、谓词、数值、单位、来源等组面,由于这些组面在检索时都需要展示给用户,这就要求把这些组面作为构成 Document 的域。从数值知识元中抽取出组成虚拟文档 Document 域的过程如图 3 所示:

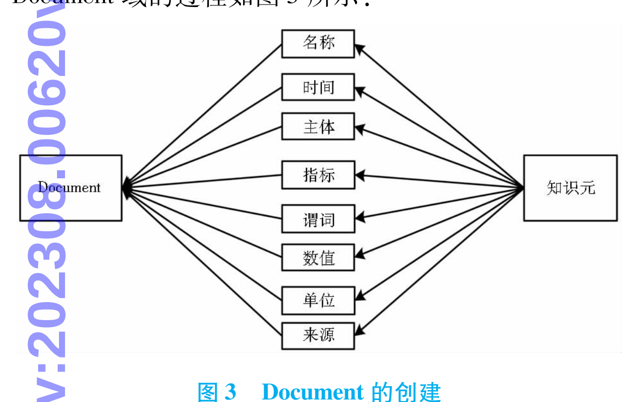


图 3　Document 的创建

需要指出的是,Document 对象中不同的 Field 域具有不同的要求和功能。为此,在从数值知识元中抽取组成 Document 的域之后,接下来要确定每个域的类型。具体包括:

(1)Keyword 域。这种类型的域不需要被分析,但是在索引过程中会被逐字地索引并存储。该类型的域比较适用于原始值,也就是那些需要被全部保留的域。在对数值知识元对象进行索引时,由于组成 Document 的“时间”域、“数值”域、“数值”域、“单位”域均不需要被分析,但是需要被索引,因此被定义为 Keyword 类型的域。

(2)UnIndexed 域。这种类型的域既不需被分析也不需要进行索引,但是该域的值需要被存储在索引文件中。该类型的域比较适合于那些需要和搜索结果一并被显示出来,但用户在检索时又不会将它的值直接用于搜索的情形。在对数值知识元对象进行索引时,由于数值知识元的“来源”域需要在检索结果中显示给用户但用户不可能用于检索,因此将“来源”域定义为 UnIndexed 类型的域。

(3)Text 域。这种类型的域需要被分析且索引,但可以被存储在索引文件中,也可以不被存储在索引文件中。在对关联数据实例对象进行索引时,作为数值知识元“名称”“主体”“指标”“谓词”部分需要被分析并且进行索引,因此将这些域定义为 Text 类型的域。

3.2.2　文本分词　本研究在对数值知识元进行中文分词和词性标注时采用了 Stanford Segmenter 中文分词器。之所以选择该分词系统的一个重要原因在于 Stanford Segmenter 中文分词器支持用户自己定义的词典,可以将自定义的词语集成到分词系统中去,从而提高分词的灵活性。数字馆藏资源中的数值知识元使用的词语通常是专业性较强的长词,因此,在对数值知识元的索引内容进行分词时,自定义一个专业词表是非常有必要的,为此,本研究定义了一个收录大量专业词汇、短语和搭配词的领域词典,以适应数值知识元的索引内容分词的需要,并将自定义的词典集成到 Stanford Segmenter 中文分词器中,从而大大提高分词的准确性。

3.2.3　索引建立　索引建立模块的主要功能是创建面向数值知识元实例对象的倒排文档,主要包括 7 个索引文件:“名称”“时间”“主体”“指标”“谓词”“数值”“单位”索引。具体实现过程如下:

(1)生成 Document。在 Document 中所有的 Field 都存储在一个 Vector 类型的数组中,以便 Lucene 遍历所有的 field 信息。具体来说,向 Document 中索引域的代码如下:

```
Document doc = new Document();
Field f1 = new Field("名称", "value1", Field.Store.YES, Field.Index.TOKENIZED);
Field f2 = new Field("时间", "value2", Field.Store.YES, Field.Index.UN_TOKENIZED);
Field f3 = new Field("主体", "value3", Field.Store.YES, Field.Index.TOKENIZED);
Field f4 = new Field("指标", "value2", Field.Store.YES, Field.Index.UN_TOKENIZED);
Field f5 = new Field("谓词", "value3", Field.Store.YES, Field.Index.TOKENIZED);
Field f6 = new Field("数值", "value2", Field.Store.YES, Field.Index.UN_TOKENIZED);
Field f7 = new Field("单位", "value2", Field.Store.YES, Field.Index.UN_TOKENIZED);
doc.add(f1);doc.add(f2);doc.add(f3);doc.add(f4);doc.add(f5);doc.add(f6);doc.add(f7);
```

黄容，何杨煜琪，王忠义，等.数字图书馆数值知识元检索系统设计[J].图书情报工作,2018,62(14):125－132.

（2）初始化 IndexWriter。初始化 IndexWriter 的主要目的是创建一个索引器。IndexWriter 索引器的主要作用是将 Document 加入到索引中去，实现面向 Document 的索引创建，并合并各种索引段，以及控制与索引相关的各方面，如删除索引等操作。

（3）创建索引。初始化 IndexWriter 之后，接下来将可以借助 IndexWriter 向索引目录中添加所有 Document。IndexWriter 提供了很多简单的接口，本研究主要借助 public void addDocument（Document doc），向索引中添加已经创建好的 Document，以实现索引的创建。

3.2.4 索引优化 优化索引主要是在建立索引之后，对整个索引目录内的索引文件进行合并，从而保证检索时的效率。为此，在完成面向关联数据实例对象的索引之后，本研究借助 IndexWriter 的 optimize（）方法对索引文件进行优化，使得索引目录中所有的索引文件合并为一个索引文件，从而大大减少目录中索引文件的数量，提高检索的速度。

3.3 数值知识元检索

3.3.1 数值知识元的名称检索 用户在输入检索词时，常常由于使用经验不足、查询处理方法的缺陷等原因，导致检索结果不能真实地反映用户的实际检索需求，查全率与查准率较低，难以形成有效的检索。由于数值知识元的名称是知识元内容中多个关键字的结合，用户在选择检索词时不一定能够精确地定位到数值知识元的完整名称，因此本研究中数值知识元的名称检索将选取模糊检索的方式。

模糊检索是通过设置单个检索词 x 在文档中的隶属度 v∈[0,1]来反馈检索结果，v 越大代表检索词与文档的相关性越高[19]。用户通过模糊检索可以改善检索结果的无序性，其检索模块会根据模糊逻辑运算得到检索结果，并按照相关度进行排序。例如，输入检索词“名称＝信息”，表示查找数值知识元库中名称包含有“信息”的所有数值知识元实体，检索结果出现的数值知识元名称可能为“中国信息产业”，也可能为“工业信息部门”等。

3.3.2 数值知识元的布尔逻辑检索 数值知识元的描述组包含主体、指标、时间、谓词、数值、单位六元组属性，为了提高查准率与查全率，在此选择布尔逻辑检索模型来构造知识描述组的检索，本研究中选择二元逻辑来进行探讨，即一系列对应于知识元特征的二元变量，包括根据各项属性从知识元库中提取出的文本检索词[20]。通过布尔逻辑检索，用户可以根据检索项在文档中的布尔逻辑关系递交查询，查询条件可以表示为由 and、not、or 等逻辑词连接的检索词序列。

4 实证

4.1 开发工具与环境

本研究采用的存储工具为 MySQL 5，开发工具为 eclipse Mar2，eclipse 用于实现知识元索引和检索功能，MySQL 用于存储知识元信息。

系统具体开发环境为：开发用 PC 机；操作系统为 win10 企业版；Java 环境为 J2EE1.7；Web 服务为 Tomcat 6.045。

4.2 系统实现

本研究从中国知网中下载了关于经济主题的相关文献，利用上文中数值知识元的识别与标引流程及方法从中抽取出数值知识元，存储并构建数值知识元库，基于数值知识元库中抽取的数据，根据上文中提出的数值知识元索引方法进行索引的构建，最终实现一个数值知识元搜索引擎。检索界面如图 4 所示：



图 4 数值知识元检索界面

4.2.1 名称检索 由于用户在选择数值知识元的名称检索词时，不一定能精确到完整的名称内容，因此在此设置模糊检索的方式来进行名称检索。例如：输入检索词“信息产业”，得到检索结果见图 5：名称中包含有“信息产业”字段的数值知识元。



图 5 名称检索结果界面

4.2.2 布尔逻辑检索 可以选择多个检索字段进行组合检索。例如，输入检索式“主体＝中国”，条件为 and，“时间＝2015 年”，条件为 or；得到检索结果为主

体为“中国”,或者时间为“2015 年”的数值知识元,如图 6 所示:

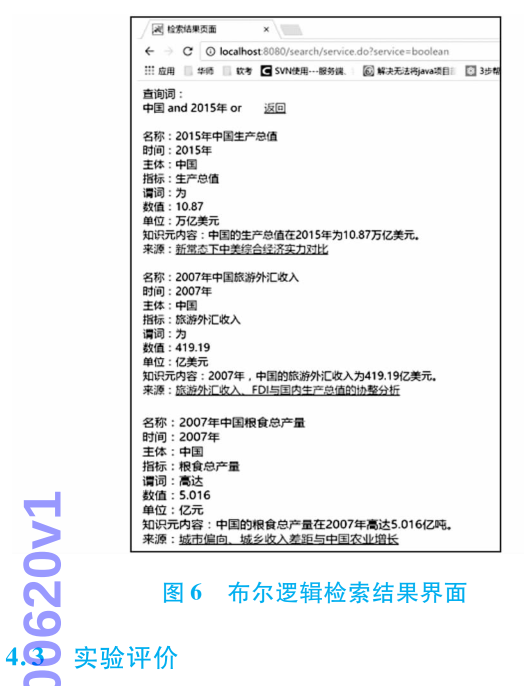


图 6　布尔逻辑检索结果界面

4.3　实验评价

为了验证数值知识元搜索引擎的有效性及实用性,本研究采用面向检索任务的主观评价方法,即通过让用户完成某个检索任务,对用户的使用体验进行分析,进而达到对数值知识元搜索引擎系统性能的综合评价。具体流程如下:

(1)设置 4 个具体的检索任务,如表 3 所示:

表 3　检索任务

任务编号	检索任务
Q_1	2015 年中国生产总值是多少?
Q_2	中国生产总值在哪一年达到了 10.87 万亿美元?
Q_3	2015 年哪个国家的生产总值达到了 10.87 万亿美元?
Q_4	2015 年中国的什么指标达到了 10.87 万亿美元?

从表 3 可以看出,4 个问句分别表示了 4 个不同的检索任务,检索任务 Q_1 是获取数值知识元的数值;检索任务 Q_2 是获取数值知识元的时间;检索任务 Q_3 是获取数值知识元的主体;检索任务 Q_4 是获取数值知识元的指标。4 个检索任务代表了数字图书馆用户不同方面的数值知识元需求。

(2)邀请 30 位数字图书馆用户(包括 15 位本科生和 15 位硕士研究生)作为实验对象,对数值知识元搜索引擎进行主观评价。

(3)挑选 3 个用户常用的知识检索工具:百度知道、CNKI、百度学术作为参照系统,评价数值知识元搜索引擎的使用效果。

(4)30 位实验对象分别借助 3 个参照系统和本研究提出的数值知识元搜索引擎完成表 3 所示的 4 个检索任务,并记录每个实验对象借助任一检索系统完成每个检索任务时点击鼠标的次数。

(5)完成检索任务后,30 位实验对象被要求立即填写表 4 所示的用户体验表。用户体验表依据李克特 5 分法将用户的满意程度分为 5 个级别:1 表示“特别不满意”、2 表示“不满意”、3 表示“一般”、4 表示“满意”、5 表示“特别满意”。30 位实验对象根据自己在完成检索任务时的使用体验进行选择。

表 4　使用体验表

	百度学术					CNKI					百度知道					数值知识元搜索引擎				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Q1																				
Q2																				
Q3																				
Q4																				

(6)依据每位实验对象的体验得分,分别计算每个检索系统在 4 个检索任务中的用户满意度得分的归一化值(用 A 表示),具体计算方法见公式(1)。其中 i 为所有检索任务中的第 i 个检索任务, Ai 表示第 i 个检索任务的用户体验得分的归一化值,j 表示所有实验对象中的第 j 位实验对象,q_{ij}表示第 j 位实验对象在完成第 i 个检索任务时的体验得分。

$$A_i = \sum_{j=1}^{30} q_{ij}/5 * 30 \qquad \text{公式(1)}$$

(7)分别计算实验对象借助每个检索系统完成每个检索任务时的平均点击次数(用 B 表示),具体计算方法见公式(2)。其中 i 为所有检索任务中的第 i 个检索任务,j 表示所有实验对象中的第 j 位实验对象,p_{ij}表示第 j 位实验对象在完成第 i 个检索任务时的点击次数。

$$B_i = \sum_{j=1}^{30} p_{ij}/30 \qquad \text{公式(2)}$$

根据实验对象对每个检索系统主观评价的体验得分的归一化值和每个检索系统的平均点击次数,对每个检索系统的性能进行定性分析,得出主观评价结果。实验结果见图 7。

图 7 展示了百度学术、CNKI、百度知道、数值知识元搜索引擎 4 个检索系统在完成知识检索任务 Q_1、Q_2、Q_3、Q_4 时的表现。

从用户体验得分来看,数值知识元搜索引擎的用户体验值得分最高,分别为 0.79、0.91、0.86、0.83,说明 30 位数字图书馆用户在完成 4 项检索任务时对数值知识元搜索引擎的使用体验最好,满意度最高。CNKI

黄容，何杨煜琪，王忠义，等.数字图书馆数值知识元检索系统设计[J].图书情报工作,2018,62(14):125－132.

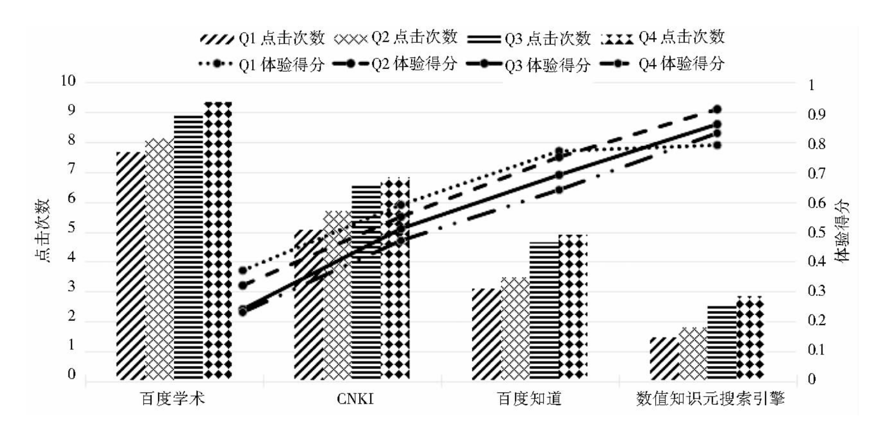


图7　实验结果

和百度知道的用户体验得分在0.5到0.7左右，说明30位数字图书馆用户在借助CNKI和百度知道完成4项检索任务时的用户满意度一般。百度学术的用户体验得分分别为0.37、0.32、0.24、0.23，得分最低，说明30位数字图书馆用户借助百度学术完成4项检索任务时的用户满意度较低。

从点击次数来看，在借助百度学术完成4项检索任务时，30位数字图书馆用户需要平均点击8次，需要的点击次数最多。在借助CNKI和百度知道完成4项检索任务时，30位数字图书馆用户需要平均点击分别为6次和4次才能检索到满意的结果，仅次于百度学术，需要的点击次数也较多。在借助数值知识元搜索引擎完成4项检索任务时，30位数字图书馆用户只需要平均点击2次便可以检索到满意的结果，需要的点击次数最少。

4项检索任务的目的是为了获取经济领域数值知识元的不同属性，之所以得到上述实验结果，通过分析发现原因主要在于：①百度学术向用户提供的是知识载体的线索，如文献题名、摘要、作者、出版年等信息，用户若想获得具体的数值知识元，需要进一步依据这些信息获取知识载体（文献），然后通过用户对文献的阅读定位查找到自己所需要的知识，这不仅意味着用户要通过多次点击才能获取自己所需要的知识，也无形中增加了用户的认知负担和成本，使得用户在完成4项检索任务时点击次数较高，用户体验得分比较低，即用户对他们的满意度较差。②CNKI与百度学术相同的是，向用户提供的仅是知识载体的线索；但与百度学术相比，CNKI的点击次数略低，用户体验值略高，原因在于CNKI作为知识检索系统不仅可以获得知识载体的线索，而且可以依据该线索信息从系统中直接获得数值知识元载体本身，然而在百度学术中，在大多数情况下，检索得到的是二次文献，若要获得知识载体本身，需要进一步链接到到其他知识检索系统，如CNKI、万方等。③百度知道是直接面向数值知识内容本身的知识检索系统，用户可以通过它们直接获得自己所需的数值知识内容本身，因此，用户只需要点击较少的次数就可以获得自己所需要的经济领域数值知识内容，这也就意味着用户使用百度知道的认知成本较低，从而获得了仅次于数值知识元搜索引擎的用户体验得分。④数值知识元搜索引擎的用户体验得分之所以高于百度知道，这是由于数值知识元搜索引擎的知识资源来源于数字图书馆，知识资源大都经过专家的评审，质量较高；而百度知道的知识资源主要来源于网络用户，不仅网络用户的知识水平参差不齐，而且知识内容本身未经第三方审核，因此知识资源的质量无法得到保障。另外，数值知识元搜索引擎也是直接面向数值知识内容本身的知识检索系统，检索入口方面具有较高的专指性，如时间、单位等，检索对象是更加细粒度的数值知识元，用户通过它们只需要点击较少的次数就可以获得自己所需要的知识，使得用户对数值知识元搜索引擎的满意度最高；而百度知道、CNKI、百度学术这些传统的检索系统入口是基于检索词（索引词）的，专指性较低，用户点击次数较高；且检索对象大多为粗粒度的文献，不能直接满足用户的知识需求，使得用户满意度相较更低。

5　结语

随着经济与科技的发展，知识作为重要的生产要素，已经成为主要的经济资源与竞争资源[21]。本研究针对数字图书馆数字资源管理的现状，以数值知识元为研究对象，提出数值知识元的识别方法与抽取流程；接着基于数值知识元的描述框架，从知识标识、知识描述、知识关系3方面提出数值知识元的标引方法与检索任务；最后构建一个数值知识元搜索引擎，实现数值知识元的检索，证明本研究提出的数值知识元标引与检索过程的可行性。随着研究的逐步深入，发现本研究还存在一些问题与不足，如没有涉及到知识元之间的链接关系研究，未能实现一个完整的数值知识元网络体系。因此，在以后的工作中，将围绕这个方面做出改进与完善，从知识元链接的角度，动态构建更为完善的知识元网络体系，实现知识元的集成化与网络化，提高数字图书馆知识服务的水平。

参考文献:

[1] 王新筠, 王海欣. 大数据背景下图书馆知识服务的思考[J]. 图书馆工作与研究, 2014(11):75-78.

[2] 赵俊娜. 高校图书馆面向科研的学科服务研究[D]. 合肥:安徽大学, 2014.

[3] 薛调. 国内图书馆学科知识服务领域演进路径、研究热点与前沿的可视化分析[J]. 图书情报工作,2012,56(15):9-14.

[4] 张海涛, 宋拓, 刘健. 高校图书馆一站式知识服务模式研究[J]. 情报科学, 2014(6):104-108.

[5] 毕崇武, 王忠义, 宋红文. 基于知识元的数字图书馆多粒度集成知识服务研究[J]. 图书情报工作, 2017,61(4):115-122.

[6] 文庭孝. 知识单元的演变及其评价研究[J]. 图书情报工作, 2007, 51(10):72-76.

[7] RONALD M. Knowledge management systems: information and communication technologies for knowledge management (third edition)[M]. Berlin:Springer,2007:265-278.

[8] 原小玲. 基于知识元的知识标引[J]. 图书馆学研究, 2007(6):45-47.

[9] 肖洪, 薛德军. 基于大规模真实文本的数值知识元挖掘研究[J]. 计算机工程与应用, 2008, 44(30):150-152.

[10] 付蕾. 知识元标引系统的设计与实现[D]. 武汉:华中师范大学, 2009.

[11] 吴超, 郑彦宁, 化柏林. 数值信息抽取研究进展综述[J]. 中国图书馆学报,2014, 40(2):107-119.

[12] 温有奎. 知识元挖掘[M]. 西安:西安电子科技大学出版社, 2005:32-35.

[13] 姜永常, 杨宏岩, 张丽波. 基于知识元的知识组织及其系统服务功能研究[J]. 情报理论与实践, 2007, 30(1):37-40.

[14] 袁阳, 肖洪. 基于知识元库自动编辑的知识服务优化[J]. 科技与出版, 2017(6):22-25.

[15] ROBERT L P, AHUJA M K. Social capital and knowledge integration in digitally enabled teams[J]. Information systems research, 2008,19(3):314-334.

[16] TAHA A. Networked library services in a research-intensive university[J]. Electronic Library, 2012(6):844-856.

[17] 化柏林. 学术论文中方法知识元的类型与描述规则研究[J]. 中国图书馆学报, 2016, 42(1):30-40.

[18] 李静. 基于知识组织的知识服务研究[D]. 天津:天津师范大学, 2008.

[19] 刘春泳. 中文问答系统中的信息检索模型的研究[D]. 重庆:重庆大学, 2007.

[20] 季拥政. 搜索引擎检索技术与检索效果探析[J]. 青海大学学报, 2006, 24(6):98-100.

[21] 德鲁克. 大变革时代的管理[M]. 赵干城,译. 上海:上海译文出版社,1999:211.

作者贡献说明:

黄容:论文架构设计与撰写;

何杨煜琪:论文校正与编排;

王忠义:论文思路设计与修改;

李春雅:实证设计及结果分析。

Research on Retrieval of Numerical Knowledge Element in Digital Library

Huang Rong[1] He Yangyuqi[1] Wang Zhongyi[1] Li Chunya[2]

[1] School of Information Management, Central China Normal University, Wuhan 430079

[2] School of Business, Nantong Institute of Technology, Nantong 226002

Abstract: [**Purpose/significance**] This paper aims to meet personalized retrieval needs of digital library users for numerical knowledge, and realize the fine-grained knowledge service. [**Method/process**] Based on the analysis of numerical knowledge element, it proposes a method of identifying, extracting, indexing and retrieving numerical knowledge elements, and constructs a retrieval system for numerical knowledge elements. [**Result/conclusion**] In addition, the case study shows that the meta-knowledge service based on numerical knowledge can improve the efficiency and user satisfaction of retrieving and using numerical knowledge.

Keywords: numerical knowledge knowledge element identification knowledge element index knowledge element retrieval